Genet. Sel. Evol. 35 (2003) 3–20
© INRA, EDP Sciences, 2003
DOI: 10.1051/gse:2002033

Original article

On the need for a control line in selection experiments: A likelihood analysis

Daniel SORENSEN*, Bernt GULDBRANDTSEN, Just JENSEN

Department of Animal Breeding and Genetics,
Danish Institute of Agricultural Sciences, PO Box 50, 8830 Tjele, Denmark

(Received 8 March 2002; accepted 5 August 2002)

Abstract – The question of whether selection experiments ought to include a control line, as opposed to investing all facilities in a single selected line, is addressed using a likelihood perspective. The consequences of using a control line are evaluated under two scenarios. In the first one, environmental trend is modeled and inferred from the data. In this case, a control line is shown to be highly beneficial in terms of the efficiency of inferences about heritability and response to selection. In the second scenario, environmental trend is not modeled. One can imagine that a previous analysis of the experimental data had lent support to this decision. It is shown that in this situation where a control line may seem superfluous, inclusion of a control line can result in minor gains in efficiency if a high selection intensity is practiced in the selected line. Further, if there is a loss, it is moderately small. The results are verified to hold under more complicated data structures *via* Monte Carlo simulation. For completeness, divergent selection designs are also reviewed, and inferences based on a conditional and full likelihood approach are contrasted.

selection / design of selection experiments / heritability / maximum likelihood estimation

1. INTRODUCTION

An old question is revisited: should selection experiments include a control line or should all facilities be devoted to a selected line only? For example, Blair and Pollak [2] gave conditions under which a control line could be avoided, if the realized heritability was inferred using mixed model methods. Thompson [15] was concerned about this recommendation and showed that the realized heritability estimator proposed by [2] was highly dependent on the prior variance (assumed known in the traditional best linear unbiased prediction approach), and little affected by the "true", unknown heritability. In view of this dependence between the inferred response and the prior variances,

* Correspondence and reprints
E-mail: Daniel.Sorensen@agrsci.dk

a formal BLUP (variances assumed known) approach does not seem to be a good proposition for analyzing selection experiments. The presence of a control line does not change this. Sorensen and Kennedy [11] proposed a two-stage "BLUP" approach, whereby variances are estimated first using a likelihood type estimator and used in a second stage to solve the mixed model equations. Concerning the need for a control line, they write: "We do not wish to imply that selection experiments should be designed without contemporaneous controls... However, if resources are severely limited, the experimenter may well wish to consider the option of eliminating the control line and to devote the facilities to selection lines, and use a mixed model approach to analyze the data." This suggestion did not build on studies of efficiency of inferences about response with and without a control line.

Those that argue for the use of control lines do so on the following grounds: (i) inferences about response to selection or heritability can be more efficient if facilities are divided into a selected and control line, rather than using all facilities in a single selected line; (ii) a control line is necessary to correct for the common environmental trend; (iii) a control line allows the response to be estimated as the difference between the mean of the selected and the control line. This is beneficial, because, as formulated by Thompson [15], it generates what Thompson refers to as an "internal evidence or test" for the inferences drawn.

In relation to (i), Hill [8] studied the conditions that justify having a control line. In contrast to the approach followed here, [8] used a model without a systematic environmental trend, but rather treated environmental effects as random variables with a zero mean and common environmental variance across generations. With this model, a control line does not necessarily lead to more efficient inferences about heritability. Hill shows that the benefit of maintaining a control line depends on the size of the common environmental variance relative to a ratio involving the phenotypic variance and the total number of individuals recorded per generation.

Regarding (ii), Hill's analysis assumed no trend in environmental effects, and he concluded that a real change in the common environment could bias heritability estimates if no control was maintained.

Our model is different from that of Hill's because we treat environmental trend as fixed effects rather than as random variables. In relation to points (i) and (ii), we show that, given the model, a maximum likelihood approach does not require a control line to infer heritability, but the control line can have a major impact on the efficiency of inferences.

The point with (iii) is that the control line may facilitate informal comparison of inferences under a variety of models. In the absence of environmental by line interactions such that common environmental effects can be assumed in the selected and control line, and if generations are discrete, it is meaningful

first to fit a simple model assuming that the mean phenotype is equal to the mean genotype, without stipulating any specific form for the genetic model. The response can then be inferred as the difference between the means of the selected and control lines. Other more highly parameterized models could then be fitted and inferences from these could be compared with those derived under the simple model. If the results are broadly similar then the operational validity of the more highly parameterized model is justified, and implementing it will lead to sharper inferences. While the idea of this informal and exploratory analysis is appealing, it may not be a simple matter to judge what can be considered as "broadly similar".

To the best of our knowledge, all three arguments above derive from calculations based on least squares estimators, using regressions or differences between phenotypic means of selected and control lines. While there are studies in the literature that compare maximum likelihood with other estimators, we are not aware of any study directly comparing efficiency of inferences about heritability or response to selection, with and without controls, using maximum likelihood. The purpose of this note is to bring together several results scattered in the literature and to focus on this problem. Specifically, the consequences of using a control line are evaluated under two scenarios. In the first one, the environmental trend is modeled and inferred from the data. In this case, a control line is shown to be highly beneficial in terms of the efficiency of inferences about heritability and response to selection. In the second scenario, the environmental trend is not modeled. One can imagine a situation in which a previous analysis of the experimental data had lent support to this decision. It is shown that in this situation where a control line may seem superfluous, minor gains in efficiency are possible if a high selection intensity can be practiced in the selected line, and if there is a loss, it is moderately small. For the sake of completeness, inferences from divergent selection experiments are also included.

The paper is organized as follows. A simple, two-generation design, in which parents are randomly mated and produce one offspring each, is studied first. This leads to simple expressions in closed form. In Section 2 the model and the likelihood are introduced and maximum likelihood estimators are derived. Section 3 derives expectations of functions of the data assuming random sampling or truncation selection. Based on these expectations, asymptotic variances of the maximum likelihood estimators are presented in Section 4. In Section 5 these results are applied to address design issues and numerical illustrations of the formulae are given. The section ends with a short discussion contrasting inferences from full and conditional likelihoods. A little simulation study involving a more complicated data structure spanning three generations, and designed to check the validity of the theoretical results, is presented in Section 6. We provide some concluding remarks in Section 7.

2. THE MODEL AND LIKELIHOOD

The data are assumed to consist of parent (generation 1) and offspring records (generation 2). In generation 1, there are $m/2$ males and $m/2$ females. The records of the males and females are assumed to have the same means and variances. Among these records, n males and n females are selected. Each male mates at random with one female, and a single offspring is produced. Thus there are n offspring records. Selection of parents takes place either at random (referred to as random selection), or on the basis of the highest scoring records (referred to as the directional selection). The former case leads to the *control line*, the latter to the *selected line*.

Asymptotic variances of heritability will be studied under two models. In the first one, it is assumed that there is one fixed effect peculiar to each generation which could represent an environmental trend. In the second model it is assumed that there is one common fixed effect across generations. This mimics the case of no environmental trend. In this second model the expression for the asymptotic variance of heritability is very intricate. A more transparent result is obtained assuming that the fixed effect is known. Numerical analyses show that this assumption has little consequences on the asymptotic variance of the estimator of heritability. The model with one common fixed effect (mimicking the absence of an environmental trend) will be labeled *model 1*, and the model with two fixed effects will be labeled *model 2*.

In the *control line* and for *model 2*, the data are assumed to be a realization from the following normal distribution:

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} \sim N\left[\begin{pmatrix}\mathbf{1}\mu_1\\ \mathbf{1}\mu_2\end{pmatrix}, \begin{pmatrix}\mathbf{V}_{11} & \mathbf{V}_{12}\\ \mathbf{V}_{21} & \mathbf{V}_{22}\end{pmatrix}\right], \tag{1}$$

where $\mathbf{y}_1$ ($\mathbf{y}_2$) is the vector of length m (n) of parental (offspring) records, μ_1 (μ_2) is the fixed effect peculiar to generation 1 (generation 2), $\mathbf{1}$ is a vector of ones of appropriate dimension, $\mathbf{V}_{11}$ is the dispersion matrix of order $m \times m$ of parental records, $\mathbf{V}_{22}$ is the dispersion matrix of order $n \times n$ of offspring records, and $\mathbf{V}_{12}$ is the dispersion matrix of order $m \times n$ of parent and offspring records. The structure of these matrices is as follows. $\mathbf{V}_{11} = \mathbf{I}\sigma^2$, where $\mathbf{I}$ is the identity matrix and σ^2 is the phenotypic variance of a single record. The element in the ith row and jth column of $\mathbf{V}_{12}$ is equal to $1/2\, h^2\sigma^2$ if i is a parent of j, and zero otherwise, where h^2 is the heritability. Finally, $\mathbf{V}_{22} = \mathbf{I}\,\sigma^2$. The model specified by (1) leads to the same form of likelihood studied by [3], from which we draw rather heavily in the first part of this note.

The implied assumption in equation (1) is that an infinitesimal genetic model is strictly operative. The heritability is twice the correlation between a parent and its offspring.

In the present scenario, response to selection is equal to $h^2\Delta$, where Δ is the phenotypic selection differential. Here we focus directly on heritability, but due to this functional relationship between response and heritability, the conclusions are equally valid for response to selection conditional on the selection differential. This parameter has been the focus of several studies in the literature (see for example, [7,8]).

The parameters of the joint distribution (1) are $\boldsymbol{\theta} = (\mu_1, \mu_2, h^2, \sigma^2)$. The joint density can be factorized as:

$$p(\mathbf{y}_1, \mathbf{y}_2; \boldsymbol{\theta}) = p(\mathbf{y}_1; \boldsymbol{\theta}) p(\mathbf{y}_2 | \mathbf{y}_1; \boldsymbol{\theta}). \tag{2}$$

The likelihood is a function of the parameter vector $\boldsymbol{\theta}$ and is proportional to (2). In view of the independence of parental records and the conditional independence of offspring records, given the parents, the likelihood can be written as:

$$L(\boldsymbol{\theta}; \mathbf{y}) = \prod_{i=1}^{m} L\left(\boldsymbol{\theta}; y_{1,i}\right) \prod_{i=1}^{n} L\left(\boldsymbol{\theta}; y_{2,i} | \mathbf{y}_{1,i}\right), \tag{3}$$

where $y_{1,i}$ is the ith record from generation 1, $y_{2,i}$ is the ith record from generation 2, and $\mathbf{y}_{1,i}$ is the vector of dimension 2 whose elements are the phenotypic values of the parents associated with $y_{2,i}$. This factorization of the likelihood has been introduced by [1] and used by, among others, [3,5]. Notice that: (i) the first term in the right hand side of (3) includes the n chosen parental records and the records from $m - n$ individuals that did not produce offspring; (ii) the second term is the contribution from the conditional likelihood of the offspring, given parental records. Since this last term is unaffected by selection on $\mathbf{y}_1$ and the first term contains all records, this has led to the conclusion that when all data on which selection is based are included in the analysis, the likelihoods with and without selection are algebraically identical and selection is said to be ignorable. A more formal proof of this result can be found in [10], and in an animal breeding context, in [9]. This means that inferences about $\boldsymbol{\theta}$ can be drawn from (3) in both the *selected line* and in the *control line*.

From (1) and (3), the likelihood is:

$$L(\boldsymbol{\theta}; \mathbf{y}) = \frac{1}{\left(2\pi\sigma^2\right)^{m/2}} \exp\left[-\frac{\sum_{i=1}^{m}\left(y_{1,i} - \mu_1\right)^2}{2\sigma^2}\right]$$
$$\frac{1}{\left(2\pi\left(1 - h^4/2\right)\sigma^2\right)^{n/2}} \exp\left[-\frac{\sum_{i=1}^{n}\left(y_{2,i} - \mu_2 - h^2\left(\bar{y}_{1,i} - \mu_1\right)\right)^2}{2\sigma^2\left(1 - h^4/2\right)}\right], \tag{4}$$

$-\infty < \mu_i < \infty$, $i = 1, 2$; $\sigma^2 > 0$; $0 \leq h^2 \leq 1$. In (4) $\bar{y}_{1,i}$ is the average phenotypic value of the parents of individual i with record $y_{2,i}$.

The likelihood (4) can be used to make inferences about $\boldsymbol{\theta}$ irrespective of whether data have been selected on the basis of $\mathbf{y}_1$ or not. However, the distribution of the maximum likelihood estimator is different with and without selection.

2.1. The maximum likelihood estimators

As in [3], the following functions of the data are defined:

$$S_1^2 = \frac{\sum_{i=1}^{m} \left(y_{1,i} - \bar{y}_1\right)^2}{m}, \tag{5}$$

the average sum of squares of generation 1, where $\bar{y}_1 = \sum_{i=1}^{m} y_{1,i}/m$, the mean of records from generation 1;

$$S_2^2 = \frac{\sum_{i=1}^{n} \left(y_{2,i} - \bar{y}_2\right)^2}{n}, \tag{6}$$

the average sum of squares of generation 2, where $\bar{y}_2 = \sum_{i=1}^{n} y_{2,i}/n$, the mean of records from generation 2;

$$S_s^2 = \frac{\sum_{i=1}^{n} \left(\bar{y}_{1,i} - \bar{y}_s\right)^2}{n}, \tag{7}$$

the average sum of squares among selected records, where $\bar{y}_s = \sum_{i=1}^{n} \bar{y}_{1,i}/n$, the mean of records from the selected parents;

$$S_{12} = \frac{\sum_{i=1}^{n} \left(y_{2,i} - \bar{y}_2\right)\left(\bar{y}_{1,i} - \bar{y}_s\right)}{n}, \tag{8}$$

which is the sample covariance between offspring records and the average records of their parents. Using these statistics, apart from a constant, the loglikelihood can be written as:

$$\begin{aligned} \ell\left(\boldsymbol{\theta}; \mathbf{y}\right) = &-\frac{m+n}{2}\ln\sigma^2 - \frac{n}{2}\ln\left(1 - h^4/2\right) - \frac{mS_1^2}{2\sigma^2} - \frac{m\left(\bar{y}_1 - \mu_1\right)^2}{2\sigma^2} \\ &- \frac{n\left(h^4 S_s^2 + S_2^2 - 2h^2 S_{12}\right)}{2\sigma^2\left(1 - h^4/2\right)} - \frac{n\left(\bar{y}_2 - \mu_2 - h^2\left(\bar{y}_s - \mu_1\right)\right)^2}{2\sigma^2\left(1 - h^4/2\right)}. \end{aligned} \tag{9}$$

Differentiating with respect to $\boldsymbol{\theta}$, setting to zero and solving yields:

$$\widehat{\mu}_1 = \bar{y}_1, \tag{10}$$

$$\widehat{\mu}_2 = \bar{y}_2 - \widehat{h^2}\left(\bar{y}_s - \bar{y}_1\right), \tag{11}$$

$$\widehat{\sigma^2} = \frac{mS_1^2 + n\left(\left(\widehat{h^2}\right)^2 S_s^2 + S_2^2 - 2\widehat{h^2} S_{12}\right)\Big/ H}{m+n}, \tag{12}$$

where $H = \left(1 - h^4/2\right)$. In (12), the contribution of the statistic S_1^2 to inferences about σ^2 is contained in the equation for $\widehat{h^2}$, which satisfies the cubic equation:

$$\left(mS_1^2/2 - nS_s^2\right)\left(\widehat{h^2}\right)^3 - (m-n)\,S_{12}\left(\widehat{h^2}\right)^2 + \widehat{h^2}\left[2\,(m+n)\,S_s^2 - m\left(S_1^2 - S_2^2\right)\right] - 2\,(m+n)\,S_{12} = 0. \quad (13)$$

These equations hold regardless of whether there is selection based on $\mathbf{y}_1$. It is clear that if a single *selected line* is used, the environmental trend can be estimated using (10) and (11). In (11) the estimator of the environmental trend at generation 2 is simply the mean of this generation minus the inferred selection response, equal to $\widehat{h^2}\left(\bar{y}_s - \bar{y}_1\right)$. A similar result in a different setting can be found in [3,5], and in a more general setting in [13]. Equations (10) and (11) show that information arising from the difference between generation means contributes to inferences about the environmental trend and not to heritability. This is not the case in the absence of an environmental trend (*model 1*).

3. EXPECTATIONS OF FUNCTIONS OF THE DATA UNDER RANDOM AND DIRECTIONAL SELECTION

Variances of the maximum likelihood estimators are here approximated using asymptotic likelihood theory. This requires computation of the expected values of functions of the data under random and directional selection.

3.1. Random selection

When the parents are randomly selected, the following expectations can be readily derived:

$$\begin{aligned}
E\left(S_1^2\right) &= (1 - 1/m)\,\sigma^2,\\
E\left(S_2^2\right) &= (1 - 1/n)\,\sigma^2,\\
E\left(\bar{y}_1 - \mu_1\right)^2 &= \frac{\sigma^2}{m},\\
E\left(\bar{y}_{1,i}\right) &= E\left(\bar{y}_s\right) = \mu_1,\\
E\left(S_s^2\right) &= \frac{\sigma^2\,(1 - 1/n)}{2},\\
E\left(S_{12}\right) &= \frac{h^2\sigma^2\,(1 - 1/n)}{2},\\
E\left[\left(\bar{y}_2 - \mu_2 - h^2\left(\bar{y}_s - \mu_1\right)\right)^2\right] &= \frac{\sigma^2\left(1 - h^4/2\right)}{n}\,.
\end{aligned}$$

3.2. Directional selection

Directional selection alters the distribution of the data. In the simple model under study, selection is on the phenotype and a proportion n out of $m/2$ are selected. This selection process will be approximated using the truncation selection model, so that $i = z/P$, and $k = i\,(i - t)$, where i represents the intensity of selection in units of standard deviations, z is the ordinate of the standardized normal curve, P is the proportion selected and t is the truncation point in standard units [4]. This is a reasonable approximation under the present assumption of normality and provided that the selected group n is of sufficient size. A similar approximation was used by [6,7,14]. Curnow [3] and Thompson [13] chose instead to treat the sums of squares involving selected records as fixed.

The following expectations can be derived:

$$
\begin{aligned}
E\left(\bar{y}_s - \mu_1\right) &= i\sigma, \\
E\left(\bar{y}_s - \mu_1\right)^2 &= (i\sigma)^2 + \frac{\sigma^2\,(1-k)}{2n}, \\
E\left(S_{12}\right) &= \frac{h^2\sigma^2\,(1-k)\,(1-1/n)}{2}, \\
E\left(S_2^2\right) &= \sigma^2\,(1-1/n)\left(1 - kh^4/2\right), \\
E\left(S_s^2\right) &= \frac{\sigma^2\,(1-k)\,(1-1/n)}{2}\,.
\end{aligned}
$$

4. VARIANCE OF THE MAXIMUM LIKELIHOOD ESTIMATOR OF HERITABILITY

4.1. A model that allows for an environmental trend (*model 2*)

Straightforward but tedious algebra, involving taking second derivatives of the loglikelihood (9) with respect to $\boldsymbol{\theta}$, computing expectations under random and directional selection using the expressions derived above and calculating the inverse of the resulting expected information matrices, leads to the following expressions for the asymptotic variance of the maximum likelihood estimators. Under random selection,

$$
Var\left(\widehat{h^2}\right) = \frac{2H}{n + \dfrac{h^4}{H}\,\dfrac{mn}{m+n}}, \tag{14}
$$

where $H\sigma^2$ is the conditional variance of offspring given the parents. Under directional selection, the variance of the *ML* estimator is:

$$Var\left(\widehat{h^2}\right) = \frac{2H}{n\,(1-k) + \dfrac{h^4}{H}\,\dfrac{mn}{m+n}}. \tag{15}$$

Since these variances depend on the parameter to be inferred, the usual likelihood approach is to substitute it by the maximum likelihood estimate $\widehat{h^2}$. In (15), in the absence of selection, setting $k = 0$ retrieves (14).

4.2. A model with one known fixed effect (*model 1*)

With one known fixed effect, under random selection, the asymptotic variance of heritability can be shown to be equal to (14). This is due to the block diagonal structure of the expected information matrix under random selection, which results in fixed effects and heritability being asymptotically uncorrelated.

With directional selection, the asymptotic variance has the following form:

$$Var\left(\widehat{h^2}\right) = \frac{2H}{n\left(1-k+2i^2\right) + \dfrac{h^4}{H}\,\dfrac{mn}{m+n}}. \tag{16}$$

With no selection, $i = k = 0$ and this expression reduces to (14). Expression (16) is smaller than (15). Note that (16) is also smaller than (14) because $2i^2 > k$.

5. DESIGN CONSIDERATIONS

Expressions (14) to (16) will be applied together with the likelihood (3) to study the efficiency of the following designs of experiments to infer heritability:

1. A single control line (*C*)
2. A single selected line (*S*)
3. Facilities divided in a selected and a control line (*SC*)
4. Facilities divided in two lines selected in opposite directions (*SS*)
5. Facilities divided in two lines selected in opposite directions and a control (*SSC*)

Model (1) predicts that response to selection in the high and low direction is symmetrical. A control line is traditionally included in a divergent selection scheme in order to test for the asymmetry of selection response. Asymmetry is

often defined as the difference in response to selection in the upward and downward direction, and is typically estimated as the difference between deviations of the high line and the control, and the low line and the control. This estimate of asymmetry requires a control line. However, other ways of studying asymmetry are possible which do not necessarily require a control line. One may wish to study the cost of including a control line (*i.e.*, efficiency of *SS versus SSC*).

In all cases, the total number of individuals scored in generation 1 is m and in generation 2 is n. These designs will be compared under *model 1* and *model 2*.

The required expressions for the asymptotic variances of heritability for cases C and S are (14) to (16). The fraction selected is $2n/m$. For design SC, it is assumed that a sample of m individuals is available in generation 1. Among these, n_c males and n_c females are randomly selected and mated to produce n_c offspring. These constitute the control line. From the remaining $m/2 - n_c$ males and $m/2 - n_c$ females, the highest scoring n_s males and n_s females are directionally selected and mated. These generate n_s offspring which represent the selected line. Here, $n_c + n_s = n$. The likelihood has three terms: the first one is like the first term in (3), and the second and third terms for the selected and control line have the form of the second term in (3). The fraction of parents chosen is $2n_c/m$ in the control line and $2n_s/(m - 2n_c)$ in the selected line. The intensity of selection is higher here than in scheme S, since $2n_s/(m - 2n_c) < 2n/m$. We will label this as SC_1.

An alternative form of the SC scheme that generates the same fraction selected as in schemes C or S, is to divide the m records of generation 1 in a selected and a control line, each of size $m/2$, and then choose n_c (n_s) as the parents of the control (selected) line. If $n_c = n_s = n/2$, the fraction selected is $2n/m$. This scheme is less efficient than the previous one. It is labeled SC_2.

For the SS design, it is assumed that from the total of m records available at generation 1, the highest scoring n_H males and n_H females and the lowest scoring n_L males and n_L females are chosen and mated ($n_H + n_L = n$). The groups generate n_H offspring (the high line) and n_L offspring (the low line). Since in this scheme we study the case where $n_H = n_L = n/2$, the proportion selected at each extreme is n/m. The SSC design is similar, except that n_H and n_L are directionally selected after n_c were randomly selected and allocated to a control line.

5.1. A model that allows for an environmental trend (*model 2*)

For *model 2*, following the same type of algebra as before, the asymptotic variance of heritability for the various designs is as follows. For the C design the variance is given by (14) and for the S design by (15). For the SC design,

the variance can be shown to be equal to:

$$Var\left(\widehat{h^2}\right) = \frac{2H}{\left(1 - k + 2i^2 p_c\right) n_s + n_c + \dfrac{h^4}{H}\dfrac{mn}{m+n}}, \tag{17}$$

where $p_c = n_c/n$, the proportion of individuals allocated to the control line. When a single selected line is used, $n_c = 0$, $n_s = n$ and (17) reduces to (15). With no selection and if a single control line is used, $i = k = 0$, $n_s = 0$, $n_c = n$, and (17) reduces to (14).

For the *SS* design the asymptotic variance is equal to (16), with the important difference that in the *SS* design, the values of i and k correspond to a proportion selected at each extreme equal to n/m, rather than $2n/m$. In common with the *C* design, the asymptotic variance for design *SS* does not depend on whether there is one known fixed effect or two unknown fixed effects. Therefore (16) holds under both *model 1* and *model 2*.

For the *SSC* design, the asymptotic variance of heritability can be shown to be equal to:

$$Var\left(\widehat{h^2}\right) = \frac{2H}{n\left(1 + \left(2i^2 - k\right)\left(1 - p_c\right)\right) + \dfrac{h^4}{H}\dfrac{mn}{m+n}}, \tag{18}$$

where $p_c = n_c/n$ has been defined in relation to (17). Expression (18) has a minimum when $p_c = 0$ where it generates (16). There is thus always an efficiency cost associated with the inclusion of a control line in a divergence selection scheme. Under the *SS* and *SSC* designs, fixed effects and heritability are asymptotically independent. Therefore (18) also holds under *model 1*.

A comparison of (17) with (15) discloses that in the model that accounts for environmental trend (*model 2*), the intensity of selection does not contribute positively to efficiency unless a control line is used. Equation (16) shows that this is in marked contrast with the model in which fixed effects are not nested within generations (*model 1*).

5.2. Model with one known fixed effect (*model 1*)

When there is a common known fixed effect in generations 1 and 2, the asymptotic variance of the heritability estimate under design *C* is given by (14), under *S* and under *SS* by (16) and under *SSC* by (18). The asymptotic variance for design *SC* can be shown to be

$$Var\left(\widehat{h^2}\right) = \frac{2H}{\left(1 - k + 2i^2\right) n_s + n_c + \dfrac{h^4}{H}\dfrac{mn}{m+n}}. \tag{19}$$

Table I. Variance of maximum likelihood estimator of heritability (multiplied by 10^2) for the 6 designs, C, S, SC_1, SC_2, SS, SSC (see text for the explanation of design symbols).

h^2	**Model**	**Design**					
		C	S	SC_1	SC_2	SS	$SSC^{(1)}$
0.1	1	0.986	0.455	0.443	0.623	0.240	0.300
0.1	2	0.986	3.108	0.692	0.880	0.240	0.300
0.3	1	0.886	0.425	0.413	0.573	0.227	0.282
0.3	2	0.886	2.447	0.634	0.795	0.227	0.282
0.5	1	0.707	0.367	0.359	0.479	0.202	0.247
0.5	2	0.707	1.591	0.525	0.643	0.202	0.247

(1) Proportion in control line equal to 0.50.

In the absence of selection, $i = k = 0$, and (19) reduces to (14). If a single selected line is used, $n_s = n$, and (19) is equal to (16).

Examples

The formulae for the asymptotic variance of heritability are evaluated numerically and the results are shown in Table I. The values of the parameters chosen are as follows. In all cases, generation 1 consists of $m/2 = 500$ males and $m/2 = 500$ females. For schemes C and S, $n = 200$ males and $n = 200$ females are selected and these generate 200 offspring (generation 2). Designs based on a selected and a control line are labeled SC, and two cases are studied, SC_1 and SC_2. For the SC_1 scheme, $n_c = 100$ males and $n_c = 100$ females are randomly selected and allocated to the control line. Of the remaining $m/2 - n_c = 400$ males and $m/2 - n_c = 400$ females, $n_s = 100$ males and $n_s = 100$ females are directionally selected and allocated to the selected line. There are $n_c = 100$ offspring measured in the control line and $n_s = 100$ offspring measured in the selected line. The proportion directionally selected is 25%.

In scheme SC_2, from the original 500 males and 500 females, half are allocated to the selected line and half to the control line. In the control line, 100 of each sex are randomly selected, whereas in the selected line, 100 of each sex are directionally selected. The proportion selected in SC_2 is $100/250 = 40\%$, smaller than in $SC1$.

In the SS scheme, from the $m/2 = 500$ individuals of each sex, the highest and lowest scoring $n_H = n_L = 100$ of each sex are directionally selected.

Finally in the SSC scheme, from the $m/2 = 500$ individuals of each sex, $n_c = 100$ from each sex were randomly selected and allocated to the control line. From the remaining $m/2 - n_c = 400$ of each sex, $n_H = 50$ of each sex

were allocated to the high line and $n_L = 50$ of each sex to the low line. In all schemes the same number of individuals are scored: $m = 1000$ in generation 1 and $n = 200$ in generation 2.

The most efficient design is *SS*, as is well known [6]. With a control line (design *SSC*) the efficiency of divergent selection decreases, as shown in the last column of Table I. However the cost of including a control line is relatively low. The figures for *SSC* in the table were generated from (18) using $p_c = 0.5$. The efficiency of designs *C*, *SS* and *SSC* is the same under *model 1* and *model 2*.

In contrast, the relative efficiency of *S versus* SC_1 and SC_2 depends on the model. If the environmental trend has to be inferred from the data, the presence of a control line leads to a very significant increase in efficiency. The variance using the *S* design is more than four times larger using the SC_1 design at low heritability values, and three times larger when heritability is intermediate.

When no environmental trend is detected (a situation mimicked by the model with one fixed effect) the control line contributes less to efficiency. If a high selection intensity in the selected line can be practised (SC_1), a minor increase can be obtained using a control line, even when only one mean is fitted in the model. When a less intense selection is practised in the selected line, (SC_2), there is a loss of efficiency of 37%, when $h^2 = 0.1$ and of 30% when $h^2 = 0.5$.

The above was obtained assuming that the common mean is known. When the mean has to be estimated the results change little. Thus, a numerical evaluation under a model where the single fixed effect is estimated, yields, for the *S* design, variances of heritability (multiplied by 100) equal to 0.507, 0.442 and 0.392 for heritability values of 0.1, 0.3 and 0.5, respectively, compared to the values in Table I: 0.455, 0.425 and 0.367.

The ranking of the designs under *model 2* agrees well with the results in [16]. These authors studied the sources of information for estimating heritability of canon bone length in sheep, in a bi-directional selection experiment spanning several generations, with a control. The amount of statistical information per individual was the largest for their *SS* design (excluding data from the control line), followed by *SSC*, *SC*, *C* and finally *S*. Their model included a year effect which makes it comparable to the results in Table I for *model 2*.

The figures in Table I for the SC_1 design were generated assuming that the same number of parents were selected in the control and selected line, such that $n_c = n_s$. In the presence of the environmental trend, differentiation of (17) shows that the optimal allocation of the *n* parents chosen is 22% in the selected line and 78% in the control line. With the optimal allocation, the variance of heritability (multiplied by 100) for values of h^2 equal to 0.1, 0.3 and 0.5 is 0.570, 0.526 and 0.443, respectively (compared to the respective figures of 0.692, 0.634 and 0.525). Figure 1 shows the variance of heritability using expression (17) and (19) as a function of the proportion chosen in the control line. The optimal proportion is little affected by the value of heritability.

Figure 1. Variance of the maximum likelihood estimator of heritability as a function of the proportion of animals allocated to the control line for the design scheme SC_1. Full line: *model 1*. Dashed line: *model 2*.

With one known fixed effect, the relationship above for the SC_1 design is rather flat. This is not the case with the model that allows for an environmental trend.

5.3. Inferences based on the conditional likelihood

The values in Table I were obtained from the full likelihood (4). A commonly used estimator is based on the conditional likelihood (here, represented by an offspring-midparent regression).

Expressions (14) and (15) are almost identical to those derived using the conditional likelihood (which involves the second term of (4) only), except that in the latter, firstly, the term $h^4mn/\left[H\left(m+n\right)\right]$ in the denominator of (14) and (15) is missing. Secondly, in a conditional likelihood analysis, only one fixed effect is identifiable. This means that when the environmental trend is modeled *via* a fixed effect peculiar to each generation, it cannot be estimated using a conditional likelihood analysis.

The extra term $h^4mn/\left[H\left(m+n\right)\right]$ originating from a full likelihood analysis, is a contribution to inference about the heritability *via* the phenotypic variance. In (14), as m goes to infinity, which is equivalent to knowing σ^2 without error, $(mn)\,/\,(m+n)$ approaches n. This is the largest possible contribution of this term for a given value of h^2. This argument builds on asymptotic properties of maximum likelihood. In fact, a little algebra shows that under random mating, the 4×4 expected information matrix associated with (4) is block diagonal, such that $(\widehat{\mu}_1, \widehat{\mu}_2)$ are asymptotically correlated, and so are $\left(\widehat{\sigma^2}, \widehat{h^2}\right)$, but the two blocks are uncorrelated with each other. Exploiting this simple structure, it is easy to show that under random mating, the asymptotic variance of $\widehat{h^2}$,

when σ^2 is known is:

$$Var\left(\widehat{h^2}|\sigma^2\right) = -\left[E\left(\frac{\partial^2 \ell\left(\boldsymbol{\theta}; \mathbf{y}|\sigma^2\right)}{\left(\partial h^2\right)^2}\right)\right]^{-1}$$
$$= \frac{2H}{n + \frac{h^4}{H}n},$$

the same result obtained *via* the limiting argument above. With selection, the expected information matrix is no longer block diagonal. However, further algebra shows that the asymptotic variance of $\widehat{h^2}$, when σ^2 is known is:

$$Var\left(\widehat{h^2}|\sigma^2\right) = \frac{2H}{n\left(1-k\right) + \frac{h^4}{H}n},$$

again, confirming that it is the phenotypic variance that indirectly contributes to the extra efficiency in the estimation of h^2 from a full likelihood analysis.

Inspection of (14) and (15) reveals that the loss of information incurred using a conditional likelihood analysis is relatively larger in the *S* design. Using the values of the parameters in the example above, the sampling variances of heritability (multiplied by 100) inferred from the conditional likelihood for heritabilities of 0.1, 0.3 and 0.5 in the *S* design are 3.200, 3.071 and 2.814, respectively. The corresponding values from the full likelihood are 0.455, 0.425 and 0.367.

In the *C* design, the loss of information is small. For the conditional likelihood at the same three heritability values, the variance (multiplied by 100) is 0.995, 0.956 and 0.875, compared to the respective values 0.986, 0.886 and 0.707 from the full likelihood.

All expressions in Section 5 hold for a conditional analysis, except for the term $h^4mn/\left[H\left(m+n\right)\right]$ which is missing in the denominators of the various formulae. Of course, as mentioned above, in contrast with a full likelihood analysis, the conditional analysis precludes correcting for the environmental trend.

Much of what has been presented can be explained by well known results from simple estimators derived from a conditional analysis. For example, the *SC* designs, and especially *SS*, capitalize on the increased variation among parents, relative to *S* or *C*, that leads to a reduction of the sampling variance of the estimator. However a comparison using *model 1* and *model 2* involving different designs based on a conditional analysis is not possible; a full likelihood approach as the one presented here is necessary instead.

Table II. Standard deviation of maximum likelihood estimator of heritability (obtained empirically from 500 Monte Carlo replicates) for designs C, S and SC_2, when the environmental trend has to be accounted for (*model 2*) or not (*model 1*).

Design	**Model**	$100\left(Var\,\widehat{h}^2\right)^{1/2}$
C	1	6.38
C	2	6.39
S	1	3.47
S	2	8.09
SC_2	1	3.86
SC_2	2	5.72

6. A SIMULATION STUDY

So far, the conclusions were based on an analysis of a simple model, and variances were obtained appealing to asymptotic results whose validity is difficult to verify. Here the results from a small simulation study are presented. The family structure is more realistic, in that parents generate several offspring (rather than only one), three generations of data are included in the analysis, and sampling variances are obtained empirically from the Monte Carlo replicates. Comparisons include only designs labeled C, S and SC_2. In all designs, a fixed number of individuals (400) are recorded per generation leading to 1 200 individuals in total.

Briefly, the designs were as follows. For C and S, 200 males and 200 females were randomly sampled and constituted generation 1. From these, 40 males and 40 females were randomly chosen (C) or selected on phenotype (S) and after random mating, five offspring of each sex were produced from each mating pair, such that 400 offspring constituted generation 2. A second round again produced 400 offspring in generation 3. For SC_2, the facilities were divided equally between the selected and the control line (100 males and 100 females in each at generation 1). From the 100 males and females, 20 were randomly chosen (C) or selected on phenotype (S), and 5 offspring of each sex were produced from each mating pair. A total of 600 individuals were available in C and 600 in S. Simulation of genotypes was based on the standard infinitesimal model. The heritability of the trait was set equal to 0.5.

As in the previous section, two models were considered. The first one, labeled *model 1*, contains 3 fixed effects, all represented in generations 1, 2 and 3. In *model 2*, there is 1 fixed effect peculiar to each generation (mimicking the environmental trend), also a total of three fixed effects. As before, *model 2* is the model accounting for the environmental trend.

The results based on 500 Monte Carlo replicates are shown in Table II. An eyeball evaluation of the sampling distribution of the *ML* estimates over the Monte Carlo replicates did not reveal signs of asymmetry. The overall picture is similar to that in Table I. In the absence of an environmental trend, *S* is the best proposition. If environmental trend has to be accounted for, excluding a control line and devoting all facilities to a single selected line increases the variance by a factor of 2 (*i.e.* $(8.09/5.72)^2$). If design SC_2 is chosen but *a posteriori* no environmental trend is detected, there is a minor loss of approximately 10% relative to *S* (3.86 *vs.* 3.47). The conclusion from this analysis agrees with the previous one: the control can be highly beneficial and it can do little damage.

7. CONCLUSION

The results available in the literature have been brought together to answer what we believe is a question not properly addressed so far. Is it worthwhile to invest facilities in a control line when inferences about heritability and derived quantities such as response to selection are based on maximum likelihood? The answer is indisputably, yes. Thus, when the model for analysis includes parameters that may represent environmental trend, a design with a single selected line leads to estimates of heritability with sampling variance that is three to four times larger than when the facilities are divided in a selected and a control line. On the contrary, if a preliminary analysis dictates that the environmental trend should not be accounted for in the model, the presence of a control line can only be slightly detrimental, and if the design had been optimized, it could even be slightly beneficial.

An important point that was not addressed in our work is the need to criticize the model(s) on which inferences are based. There is a vast literature on this topic from classical and Bayesian perspectives, and this is not the place to discuss this important subject which is the focus of much current research. However, as pointed out in the Introduction, the presence of a control line can allow informal comparison of inferences under a variety of models. Control lines will probably also contribute to sharper conclusions using more formal methods of model comparisons, although we are not aware of specific studies addressing optimization of designs for model comparisons.

The results presented here show that the way information is used to infer heritability *via* maximum likelihood in an experiment involving a selected and a control line, depends critically on whether fixed effects are nested or cross-classified with generations. A similar result in the context of prediction of breeding values was obtained by [12]. This is in contrast with a divergent selection scheme, with or without a control line, where efficiency is not dependent on the distribution of fixed effects across generations.

The central message of this paper can be summarized invoking the old Russian motto: trust is good, control is better.

REFERENCES

[1] Anderson T.W., Maximum likelihood estimation for the multivariate normal distribution when some observations are missing, J. Amer. Stat. Assoc. 52 (1957) 200–203.
[2] Blair H., Pollak E.J., Estimation of genetic trend in a selected population with and without the use of a control population, J. Anim. Sci. 58 (1984) 878–886.
[3] Curnow R.N., The estimation of repeatability and heritability from records subject to culling, Biometrics 17 (1961) 553–566.
[4] Falconer D.S., Introduction to Quantitative Genetics, 4th edn., Longman, New York, 1996.
[5] Henderson C.R., Kempthorne O., Searle S.R., von Krosigk C.M., The estimation of genetic and environmental trends from records subject to culling, Biometrics 13 (1959) 192–218.
[6] Hill W.G., Design of experiments to estimate heritability by regression of offspring on selected parents, Biometrics 26 (1970) 566–571.
[7] Hill W.G., Design and efficiency of selection experiments for estimating genetic parameters, Biometrics 27 (1971) 293–311.
[8] Hill W.G., Estimation of realised heritabilities from selection experiments. II. Selection in one direction, Biometrics 28 (1972) 767–780.
[9] Im S., Fernando R.L., Gianola D., Likelihood inferences in animal breeding under selection: a missing-data theory view point, Genet. Sel. Evol. 21 (1989) 399–414.
[10] Rubin D.B., Inference and missing data, Biometrika 63 (1976) 581–592.
[11] Sorensen D., Kennedy B., Analysis of selection experiments using mixed model methodology, J. Anim. Sci. 63 (1986) 245–258.
[12] Sorensen D., Johansson K., Estimation of direct and correlated responses to selection using univariate animal models, J. Anim. Sci. 70 (1992) 2038–2044.
[13] Thompson R., The estimation of variance and covariance components with an application when records are subject to culling, Biometrics 29 (1973) 527–550.
[14] Thompson R., Design of experiments to estimate heritability when observations are available on parents and offspring, Biometrics 32 (1976) 283–304.
[15] Thompson R., Estimation of heritability in a selected population using mixed model methods, Génét. Sél. Évol. 18 (1986) 475–484.
[16] Thompson R., Atkins K.D., Sources of information for estimating heritability from selection experiments, Genet. Res. 63 (1994) 49–55.